DOCUMENT RESUME

ED 366 633 TM 021 031

AUTHOR Burry-Stock, Judith A.; Oxford, Rebecca L.
TITLE Expert Science Teaching Educational Evaluation Model (ESTEEM) for Measuring Excellence in Science Teaching for Professional Development.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC. Office of Research.
PUB DATE [93]
CONTRACT R117Q00047
NOTE 39p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; Constructivism (Learning); *Criteria; Educational Assessment; Educational Change; Educational Practices; Educational Quality; Elementary Education; Elementary School Students; *Elementary School Teachers; *Evaluation Methods; Models; Outcomes of Education; *Professional Development; Qualitative Research; Science Education; *Science Instruction; Thinking Skills
IDENTIFIERS Expert Novice Problem Solving; Experts; *Expert Science Teaching Evaluation Model; *Teacher Centered Instruction

ABSTRACT

This paper discusses the development and theoretical premises of the Expert Science Teaching Evaluation Model (ESTEEM) as well as research into this model from 1990 through 1993. ESTEEM was developed under the aegis of the Center for Research on Educational Accountability and Teacher Evaluation (CREATE) with major goals of defining expert science teaching, developing instruments to assess it, and developing an expert teaching model. Both qualitative and quantitative analyses have been done to develop and validate the first instruments and have involved nearly 200 fourth- through eighth-grade teachers. The studies have indicated that expert teachers are defined by the evaluation criteria and that there is only about a 50% agreement in the top quartile of teachers sorted by two developed measures, the Classroom Observation Rubric and the Student Outcome Assessment Rubric, suggesting caution in considering what expert means. The theoretical base derives from the constructivist movement, Scriven's Duty Based Evaluation, and the novice through expert literature. Research so far suggests that most science teaching is done from a teacher-centered approach, disseminating content at the lower end of the higher order thinking skills continuum. Two figures and 14 tables present some research findings. (Contains 73 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 366 633

EXPERT SCIENCE TEACHING EDUCATIONAL EVALUATION MODEL (ESTEEM) FOR MEASURING EXCELLENCE IN SCIENCE TEACHING FOR PROFESSIONAL DEVELOPMENT

Judith A. Burry-Stock
Rebecca L. Oxford
The University of Alabama

This article has been accepted for publication in a special edition of the *Journal of Personnel Evaluation in Education* in a special edition on the Center for Research and Educational Accountability and Teacher Evaluation (CREATE)

This work is being field tested as part of a funded project by the Center for Research on Educational Accountability and Teacher Evaluation (CREATE).

This work was supported by the Office of Research, Office of Educational Research and Improvement, U.S. Department of Education (Grant No. R117Q00047). The opinions expressed are those of the author(s), and no official support by the U.S. Department of Education is intended or should be inferred.

TM021031

BEST COPY AVAILABLE

EXPERT SCIENCE TEACHING EDUCATIONAL EVALUATION MODEL (ESTEEM) FOR MEASURING EXCELLENCE IN SCIENCE TEACHING FOR PROFESSIONAL DEVELOPMENT

Judith A. Burry-Stock
Rebecca L. Oxford
The University of Alabama

What is excellence in science teaching? How does one become a science expert? Why does one choose to become an expert? How long does it take to become an expert in science teaching? How does one stay motivated to reach a satisfactory level of expertise in science teaching? How does one know his/her level of expertise?

These are some of the key questions addressed via the Expert Science Teaching Evaluation Model (ESTEEM) under the aegis of the Center for Research on Educational Accountability and Teacher Evaluation (CREATE), funded by the U.S. Office of Educational Research and Improvement (OERI). This model, its development, theoretical premise, and research from 1990 through 1993, are the focuses of this article.

The major goals of the CREATE Expert Science Teaching Project which developed the ESTEEM are to define characteristics of expert science teaching, develop instruments to assess expert science teaching based on these features, and develop an expert teaching evaluation model to improve science instruction. The project involved three years of study and is still under development. Both quantitative and qualitative data analyses were done to develop and validate the first instruments. Nearly 200 fourth- through eighth-grade teachers from seven states were involved during the three years 1990-1993.

Related Literature

Science education is often overlooked as being important in the lives of today's children. Yet, the knowledge is critical for students today, as it is the study of how things work. Science requires that students acquire knowledge by careful observations, by deducing laws that regulate changes and conditions, and by testing these conclusions. Knowing how to do these things is important to survival and should be an important part of the school curriculum. Many teachers consider science complicated and difficult to teach. However, in our technological world, it is more important than ever that youth understand the basic fundamentals.

Often the vocabulary of the various sciences is so overwhelming to students that they fail to grasp even simple ideas. For students to really learn science, they must be able to understand the concepts. Conceptual understanding goes beyond the memorization of facts and definitions, and it may require a special kind of teaching. A science teacher needs to be not only knowledgeable about science, but also able to provide a stimulating environment for students that facilitates the understanding of concepts. Understanding the context of the teaching/learning environment is critical to the development of expertise in science teaching.

Evertson (1990, p. 3) offered a model (shown in Figure 1) for investigating science teaching and learning. The "instructional event" brings teachers, students, and materials together to create a learning environment, according to Evertson. This happens in the context of the school and community and in the still wider context of the state, region, and nation. If the teacher is not familiar with all aspects of this process he or she cannot provide a relevant and meaningful learning experience. Likewise, if the learning context is not considered in the evaluation of teaching and learning, then the evaluation loses its point of reference.

The following literature reviews topics that are inherent to the development of the ESTEEM. Topics such as evaluation, educational reform, teacher evaluation, expert versus novice teaching, and constructivism provide the foundation of the theoretical perspective for the development of the instruments and the professional development aspects of the ESTEEM.

Evaluation's Driving Role in Curriculum

One of the most powerful processes affecting our schools is evaluation. *Educational evaluation* is the measurement of value or worth of any aspect of education. By contrast, *educational assessment* gauges how things stand--the status of education--without adding any value judgments. Assessment suggests differentiating among performances (Hopkins, Stanley, & Hopkins, 1990), which requires measurement but does not necessarily demand a real decision about value or worth. In practice, however, it is almost impossible to consider the status or conditions of education without making judgments about worth or value, so evaluation and assessment are frequently used as synonyms.

Evaluation often drives education, particularly the curriculum. Criteria for evaluation frequently force educators to organize learning to produce intended outcomes. Educators' beliefs about the power of test results affect participants (Madaus, 1987). If test results are to lead to important decisions, then teachers "will teach to the test." Over a period of time, evaluation practice thus begins to define the curriculum. In the Tennessee STAR project, a study of class size, Evertson and Randolph (in press) found that teachers taught to the state testing objectives, and that the emphasis was on the products of learning, not the process of learning.

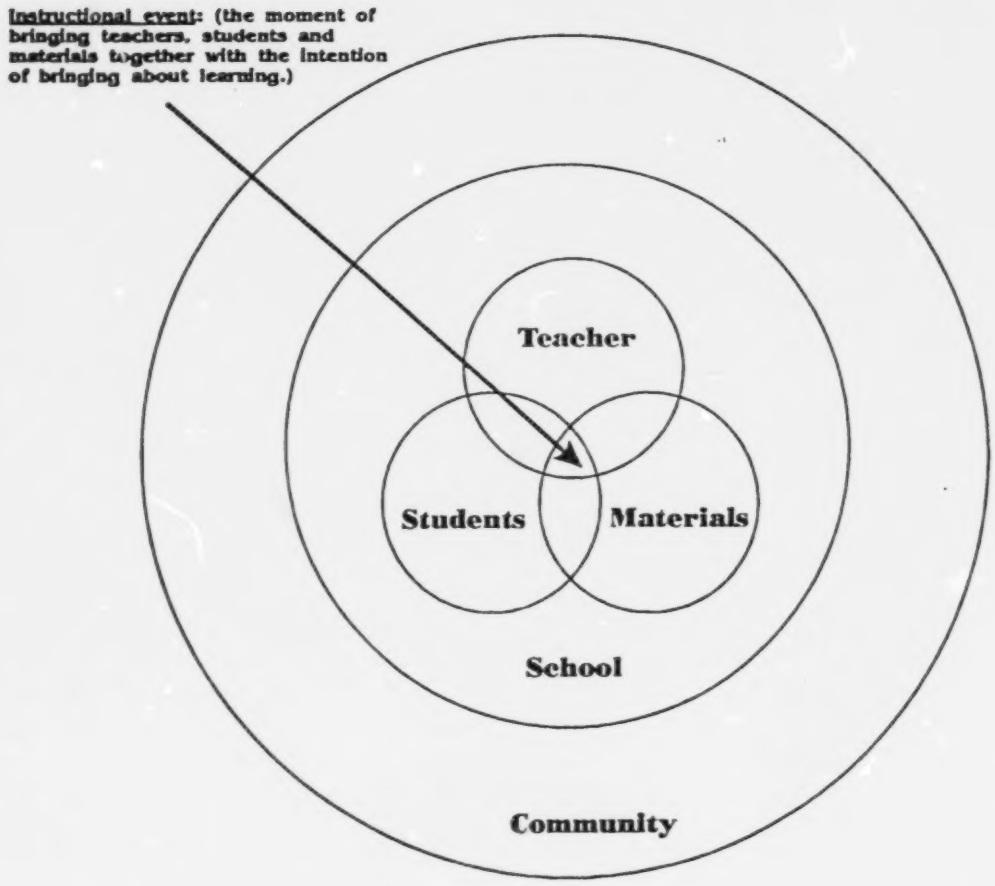

Figure 1. Model for investigating science teaching and learning (Evertson, 1990, p. 3).

Teacher evaluation often helps propel a curriculum in a certain direction. Teacher evaluation is conducted for entrance to professional teacher education programs and for licensure, certification, promotion, tenure, and professional development (Dwyer, 1991; National Board for Professional Teaching Standards, 1991). If teachers are evaluated on a certain set of standards, they are likely to teach to the evaluation model and become product-oriented rather than process-oriented (unless of course the evaluation model itself is geared toward process).

If major changes are to occur in teaching and learning, then educational evaluation models must change. To make changes in teaching and learning, we need to understand what teaching and learning are, and we also need to comprehend how any changes in teaching and learning practices are related to alterations in evaluation procedures. We need to tie these understandings to educational reform efforts.

Educational Reforms

The latest round of educational reform, from the 1970s to the 1980s, was characterized by growing disenchantment with our schools, which produced an assortment of initiatives to restore the American schools to a position of prominence (Murphy, 1990a). Most of these initiatives, labeled as the "standard-raising movement," were used to implement change through a top-down approach. Top-down refers to a reform effort at the national and state-wide levels rather than local, bottom-up, administrative control.

However, the failure of the top-down approach has given rise to hopes that a bottom-up strategy will be more successful (Murphy, 1990b, 1991). This bottom-up approach is reiterated in the so-called fourth generation evaluation models (Guba & Lincoln, 1989) and in the restructuring of schools (Evertson & Murphy, in press). Suggestions for restructuring American schools include, among others: (1) greater empowerment and professionalism of teachers; (2) a shift to student-centered learning rather than teacher-centered learning; (3) the use of rational, well organized curricula; and (4) more flexible classroom organizational structures (Bliss, Firestone, & Richards, 1991; David, 1989; Elmore, 1989; Murphy, 1991).

The teaching/learning environment of schools must be changed in this country (Barth, 1988; Beck & Murphy, in press; Marshall, 1988, 1990; McCarthy & Peterson, 1989; Wise, 1989). Kyle (1991) stated that the needed reform must involve more than surface change. What is taught, why it is taught, and how it is taught must reflect the best in educational research. Current reform efforts not only seek to make changes in scope and sequence, but also seek "to restructure the overall educational system to make it responsive to the needs of *lifelong* learners in the information age" (Salinger, 1991, p. 30).

Teachers have a significant role in this restructuring. To have new forms of school management and governance, we need to redefine the teaching profession (Evertson & Murphy, in press) and encourage teachers to teach for greater student understanding (Marshall, 1988, 1990). Teachers will need to interact with colleagues, participate in school decisions, and assume leadership roles. Teachers will need to become more interdependent with their peers and to share knowledge. Teachers need to develop organizational structures that "break down traditional teacher isolation in the classroom" (p. 11). They must be more concerned with the purposes of education than with achieving isolated goals (Conway & Jacobson, 1990; Petrie, 1990). "A professional work culture that fosters teaching for meaningful understanding will help support the development of organizational structures" (Evertson & Murphy, in press, p. 309).

Teacher Evaluation

Teacher evaluation has an important new role that goes far beyond mere accountability. This new role is to promote professional growth by clearly delineating what "expertness" or "effectiveness" means in teaching, by determining how close a given teacher comes to these criteria, and by providing diagnostic information to assist in the teacher's growth toward the target characteristics. If the evaluation criteria do not reflect expertness, then teachers will not develop high levels of performance, and their students will be less likely to do their best.

Educational evaluators have spent considerable energy identifying the differences between expert and novice teachers. Comparisons of novice and expert teachers' interpretations of classroom events indicate that experts have deeper, more richly connected knowledge structures (*schemata*) to draw upon when making a decision. Novices have leaner, less developed *schemata*, presumably because of lack of experience (Leinhardt & Greeno, 1986). Well-developed *schemata* allow the expert teacher to first determine which classroom events merit attention and then immediately to pull other relevant information from memory to decide an appropriate response (Carter et al., 1988). Thus, a key factor in expert teachers' thinking is automaticity of response.

In the expert-novice comparison by Borko and Livingston (1989), expert teachers improvised their actions from sketchy lesson plans when students needed change. The experts had scripts (both content and process) stored in memory and were able to access them quickly. Thus, expertise is characterized by what appears to be a smooth, flowing, and automatic performance (Dreyfus & Dreyfus, 1986). Novices, lacking such *schemata* and the automaticity to go with them, are not able to make a new meaningful plan on the spot. Researchers disagree on the degree to which an expert teacher is able: (1) to apply generalized *schemata* to a problem; or (2) to construct new solutions on the spot, based on the context of a particular problem. Using either approach, the responses are anchored in teachers' experiences. Expert teachers also engage in more self-regulated, purposeful behavior than do novices (Sparks-Langer & Colton, 1991, based on the results of Leinhardt & Greeno, 1986). This behavior is known as *metacognition*.

Effective teachers show a practical knowledge of their craft, the "wisdom of practice" (Leinhardt, 1990, p. 18). They have a sound repertoire of information that weak and inexperienced teachers do not have. Because "experienced teachers do not normally pass on their craft in an explicit way, most teachers have not had to separate unverified opinion from probable truth" (Ball, 1988, p. 19). Though it is difficult to separate teacher-provided information into important and valid vs. unimportant and nonvalid, it must and can be done, according to Leinhardt. This can be accomplished by: (1) identifying expert teachers; (2) obtaining detailed descriptions of their teaching practices; and (3) distilling information by some shared and publicly visible system based on cognitive psychology or anthropology. Patterns of successful teaching can be identified through approaches that deal with information-processing strategies, memory structures, and the contexts of teacher behavior.

Recently, researchers have tried to go beyond merely contrasting expert and novice teacher behavior. One current focus is identifying teacher education activities that promote reflective thought in beginning teachers, and another focus is finding out how teachers can progress from the novice state to the expert state.

Course activities designed to promote reflective thinking helped students develop *schemata* that were more like those of experts (Morine-Dershimer, 1989). In a longitudinal study of a teacher education program, Hollingsworth (1990) found little change in pre-service teachers' concern with student learning as opposed to technical aspects of teaching until the teachers' second or third year of actual teaching. At that point, when their scripts for everyday management and instructional activities became automatic, these teachers could begin to focus on student outcomes more fully.

One lesson that we might draw from research is that we should teach novice teachers the *schemata* of expert teachers. However, acting on this conclusion would subvert the lessons learned from "constructivism" (each of us must construct our own meaning) and from "situated cognition" (expert teachers draw upon their own knowledge and prior experience to develop their own wisdom of practice).

Trying to teach novices the *schemata* of experts would also short-circuit the development of professional judgment. Research can tell us how complex and uncertain teaching is, but it cannot describe the kinds of decisions teachers must make in any particular situation (Lambert & Clark, 1990).

Constructivism

As noted earlier, the overall goal of CREATE's expert science teaching project is to improve science teaching. Most of the informed professionals in the science teaching field today espouse constructivism as the most effective way to teach for meaningful understanding. The constructivist model assumes that students have a purpose for learning, and that they are actively engaged in constructing meanings from their learning experiences. Yager (1991) called constructivism the most exciting idea of the past 50 years, and suggested the model may serve as a link connecting all current lines of research in science education.

Constructivism is an ontological/epistemological paradigm addressing both what is known and how it is known. Ontology refers to the area of metaphysics concerned with the core of things. Epistemology is a branch of philosophy that pertains to the study of the origin, foundation, limits, and validity of knowledge. In simple words it means the study of the way of knowing. Prominent constructivist Ernst von Glasersfeld traced the theory of cognitive construction to Neapolitan philosopher Giambartista Vico, who wrote in 1710 that individuals know only the cognitive structures that they put together themselves (Yager, 1991). Vico explained that "to know" means "to know how to make" and that a person knows "a thing" only when he or she can explain it for others to understand and use. "Constructivists do not consider knowledge to be an objective representation of an observer-independent world. For them, knowledge refers to conceptual structures that epistemic agents consider visible" (Yager, p. 54). Individuals who promote a constructivist perspective believe that each learner creates his/her unique understanding.

Constructivism deals not only with how understanding is constructed but also how the structure of cognition affects behavior. The theory of constructivism suggests that individuals interpret and act according to conceptual categories in the cognitive system. This theory is like many other cognitive approaches in explaining that an event does not just present itself to the individual in raw form; instead, the person constructs experience according to the organization of

the cognitive system. Constructivism assumes that individuals are creative and dynamic. Instead of merely being acted on by their situations, people act dynamically to affect changes based on how they think and regulate their activities.

Constructivism is partly based on Kelly's (1955) theory of personal constructs. He suggested that people understand experience by grouping events according to similarities. A construct is a distinction between opposites, such as short-tall, fat-thin, black-white, that is used to understand events, things, and people. A person's cognitive system consists of numerous such distinctions. By classifying an experience into categories, the person gives meaning to the experience.

Constructs are organized into interpretive schemes, which identify what something is and place the object or event into a category. With interpretive schemes, we make sense out of something by putting it into a larger context of meanings. Interpretive schemes are developed during the maturation process according to the orthogenetic principle, by moving from relative simplicity and generality to relative complexity and specificity (Werner, 1957). Thus, very young children have simple construct systems, and adults have more complicated ones.

Constructivism recognizes that constructs have social origins; they are learned through interaction with others. An individual's construct system is a direct result of interaction in social groups and cannot be separated from social life. Goal achievement is often a social phenomenon involving the coordination of goal-oriented strategies by a number of people. Culture is particularly relevant in determining the meaning of events, people, and things (Applegate & Sypher, 1988) and in helping coordinate goal-oriented efforts. If a group of people share attributions and perceptions, they are more likely to behave similarly than if they do not share attributions and perceptions. Any moment of interpersonal behavior reflects each person's history and cognitive understandings. Behavior follows, at least to some extent, *schemata* that allow people to interact because they know the behavior is appropriate to the circumstances.

Although constructivism recognizes social interaction and culture, it remains primarily a psychological and cognitive theory dealing with individual differences in construct complexity. The idea of cognitive complexity was originally developed by Crockett (1965) and elaborated by Schroder, Driver, & Streufert (1967). Individuals with highly developed interpretive schemes make more discriminations than those who see the world simplistically. Adults differ widely in their cognitive (construct) complexity. Different parts of the construct system of a single person can also differ in complexity, so an individual might have very elaborate, complicated constructs about science but not about art (or vice versa). Complexity or simplicity in the cognitive system is a function of the relative number of constructs and the degree of differentiation one can make between the elements of experience. The relative number of constructs used by an individual to organize a perceptual field associated with "cognitive differentiation."

Cognitive differentiation affects the number of goals one can achieve through action. Often an action involves multiple intentions and may embody several strategies at the same time. Goal achievement is often attached to one's own desires and cognitively differentiated perceptions.

Interpersonal constructs are especially important because they guide how we understand other people. Individuals differ in the complexity with which they view others. Cognitive simplicity leads to stereotyping other people, whereas differentiation allows more subtle and sensitive distinctions to be made and permits greater understanding of others' perspectives.

Pedagogy has been dominated by a behavioristic approach to teaching, based on the traditional belief in an objective reality. Behaviorists perceive teaching and learning as behaviors. The teacher presents a small set of stimuli and reinforcers to get students to emit an appropriate behavioral response. However, if the goal is for students to understand, conceptualize, and apply new information, behaviorism is not successful, because there is no model for understanding. Constructivism is a far more effective approach.

Von Glaserfeld (1988) argued that we should think of knowledge as cognitive mapping of what turns out to be workable and not a representation of what exists. If this is so, then curriculum materials and instruments should be designed more effectively. Teachers should realize that rote learning and repetition do not generate understanding and useful knowledge.

Group learning, where pairs and small groups of students solve problems, is one of the key constructivist techniques. Constructivists engage students in problem solving and higher-order thinking skills such as those described by Resnick (1987). Teaching procedures that illustrate constructivism were highlighted by Yager (1991):

Seeking out and using student questions and ideas to guide lessons and whole instructional units;

Accepting and encouraging student initiation of ideas;

Promoting student leadership, collaboration, location of information, and taking actions as a result of the learning process;

Using student thinking, experiences, and interests to drive lessons (this means frequently altering teachers' plans);

Encouraging the use of alternative sources for information both from written materials and experts;

Using open-ended questions and encouraging students to elaborate on their questions and their responses;

Encouraging students to suggest causes for events and situations, and encouraging them to predict consequences;

Encouraging students to test their own ideas, i.e., answering their questions, their guesses as to causes, and their predictions of certain consequences;

Seeking out student ideas before presenting teacher ideas or before studying ideas from textbooks or other sources;

Encouraging students to challenge each other's conceptualizations and ideas;

Using cooperative learning strategies that emphasize collaboration, respect individuality, and use division of labor tactics;

Encouraging adequate time for reflection and analysis; respecting and using all ideas that students generate; and

Encouraging self-analysis, collection of real evidence to support ideas, and reformulation of ideas in light of new experiences and evidence. (p. 55-56)

Novice to Expert

Berliner (1987) and Dreyfus and Dreyfus (1986) approached teaching developmentally as a theory of skill learning with the following stages: (1) novice; (2) advanced beginner; (3) competent; (4) proficient; and (5) expert. Teacher education (aided by assessment) should provide the framework for moving the teacher from the novice stage to the advanced beginner stage and ultimately the expert stage.

Dreyfus and Dreyfus (1986) delineated the five steps named above by studying artificial intelligence and the problem-solving capability of computer programmers. Teacher behavior is considerably more complicated and multi-faceted than the problem-solving activities of the computer programmers. The measurement of classroom behavior is a complex task.

The *novice stage* is characterized by skill development. The novice learns to recognize various features and facts and to determine rules. Beginning teachers who want to do a good job lack any rational sense of overall context or organization and judge their own performance by how well they follow learned rules. If student discipline problems occur, beginning teachers do not have the experience to be flexible with rules.

The *advanced beginner stage* is characterized by the importance of broad skills and by the use of more sophisticated rules. Because the advanced beginner can detect similarities with prior examples, the elements are called "situational" to differentiate them from context-free elements. Advanced beginning teachers may now determine in which situation one set or one rule works differently from another. The situation determines how the discipline rules are to be applied.

The *competent teacher stage* is exemplified by teachers who cope with problems and students in a hierarchical process of decision-making. This is characterized by first choosing a plan to organize the situation, and then identifying a small set of factors that will help improve the situation. Generally, a competent performer has a goal in mind and sees a new situation as a set of facts. The competent teacher now has sufficient experience to determine when the classroom rules will work, and when the situation requires an entirely different set of procedures not covered by a set of rules. The teacher feels a sense of personal responsibility. In the discipline example above, the competent teacher would consciously choose the rules and goals based upon the situation, but would feel a sense of personal responsibility for the outcome.

The *proficient teacher stage* involves thinking analytically, but intuitively organizing and understanding the task. The performers' experience is the background for determining how best to manipulate the environment. Involvement in a specific skill is no longer of primary importance. The proficient teacher recognizes a very large repertoire of patterns. Proficient teachers experiencing a discipline problem do not make decisions based upon rules. They examine their experiences, deliberately consider the alternatives, and feel a sense of responsibility for the outcome.

The *expert teacher stage* is established on maturity and practical understanding. They do not stop to deliberate when making certain decisions, but instead they automatically and fluently perform. Experts are deeply involved in coping with their environment and do not see problems in a detached way. They do not work at solving problems nor worry about the future, conceive plans, or verbalize issues. Fluid performance is characteristic of experts. Experts do not take the time to think, but they know by feel and familiarity what action to take. A champion basketball player does not stop to analyze how to make the basket. The decision to shoot for a basket is intuitive. A chess master can roughly recognize 50,000 types of positions. Expert teachers with the discipline problem know intuitively what to do.

Culmination

The literature reviews are summarized nicely by the National Center for Improving Science Education, which studied ways to devise appropriate curriculum, instruction, assessment, and teacher development and support (Loucks-Horsley et al., 1990.) These findings can help to guide us in solving the many science education problems. See Table 1 for the recommendations of the National Center's study.

Table 1

Findings of the National Center for Improving Science Education (Loucks-Horsley et al., 1990, p. xiii)

1. Make science a basic.

Curriculum: What Should We Teach?

2. Build curricula that nurture conceptual understanding.
3. Connect science attitudes and skills as important goals.
4. Include scientific attitudes and skills as important goals.

Instruction: How Should We Teach?

5. View science learning from a constructivist perspective.
6. Use a constructivist-oriented instructional model to guide learning.

Assessment: How Can We Identify Successful Learning?

7. Assess what is valued.
8. Connect curriculum, instruction, and assessment.
9. Use a variety of assessment strategies.

Teacher Development and Support: How Can We Prepare and Support Teachers to Teach Science Well?

10. Assess programs as well as students.
11. View teacher development as a continuous process.
12. Choose effective approaches to staff development.
13. Provide teachers with adequate support to implement good science programs.

Teacher Performance Assessment Assumptions

In order to assess the validity of teacher evaluation instruments, the following assumptions were adopted. The following assumptions based upon the current research, just reviewed, are inherent in the development and implementation of this project. These assumptions synthesize the above literature review on which our project is built (Stufflebeam, 1990).

1. Pedagogical knowledge is developmental.
2. Valid instrumentation is built upon a research base, an empirical base, or a combination of the two.
3. Teaching is multifaceted; therefore, it must be assessed through multiple methods of data collection.
4. Teacher performance assessment behaviors may be both generic and content-specific depending upon the stage of the teacher's professional development and the context of the evaluation.
5. Measurement purposes are established a priori.
6. Measurement instruments and procedures vary depending upon the inference(s) that is to be made from the assessment outcomes.
7. Validity, reliability, inter-rater agreement, and utility of the inference are confirmed post hoc.
8. Teacher performance assessment is related to the context, the classroom, building, and community.
9. Observers are well trained in the process of observing.
10. Observers fully understand the meaning(s) of the teaching behaviors to be observed.
11. Observers are well trained to implement the teacher evaluation standards.

Design of the ESTEEM

The findings of the National Center for Improving Science Education detailed above provided a general framework for the ESTEEM assessment rubrics. In addition to these conclusions, the ESTEEM also takes into consideration the draft of the *National Science Education Standards: An Enhanced Sampler* published by the National Committee on Science Education Standards and Assessment from the National Research Council in 1993, as well as the criteria to be used for assessing science teachers through the National Board of Professional Teaching Standards (1992). Over 500 literature references annotated for the CREATE Expert Science Teaching Project (Brennan, Zhang, Slater, & Bolland, 1993) guided the development of the ESTEEM and the ESTEEM instruments and rubrics. The assessment instruments currently available provide information on five aspects of expert science teaching (teacher and student behaviors): (1) a classroom observation of teaching and student behaviors; (2) recall and conceptual student outcomes for one lesson; (3) a recall and conceptual student outcome rubric to be used at the end of a unit of study; (4) a teacher's self-report of teaching practices; (5) and a teacher's self-report of his/her grading practices. A constructivist, student-centered, perspective underlies the ESTEEM. However, the classroom observation rubric also strongly reflects expert science teaching.

The ESTEEM teacher evaluation instruments are also rooted in the philosophy that it is the teacher's duty to provide the best process and environment for enhancing students' learning. Michael Scriven's "duties-based" model for teacher evaluation (1991) includes five categories in the ESTEEM model: (1) subject matter knowledge; (2) skills in teaching; (3) skills in assessment; (4) professionalism; and (5) other contributions to the school. This highly useful model was adopted by the ESTEEM and included in the teaching practices addressed by the instruments. It serves as the basis for several ways to evaluate science teachers: peer-evaluation; outside evaluator; and self-evaluation.

The ESTEEM teacher evaluation instruments view expert science teaching as specific kinds of teaching practices, as opposed to a person or a product. Thus, the instruments in the model must measure science teaching performance and related factors from multiple perspectives in order to capture the multifaceted environment of the classroom (Burry, in press). The task of evaluating expert science teaching is an enormous and overwhelming undertaking. The five instruments described in this manuscript have taken over three years to develop and pilot. There are two more instruments in draft form: instructional design and reflective teaching practices. We intended these instruments that comprise the ESTEEM to be used as a source for professional development. They may be implemented at different stages of a teacher's career from novice to expert. We recommend that one or two instruments be used at a time. The instruments should be used over a several-year period as it may take more than a year to change teaching practices and student behaviors (Burry, Oxford, & Bolland, 1993).

Research and Development of the ESTEEM Instruments

A Picture of Expert Science Teaching: At the time of this writing the ESTEEM includes five science teaching evaluation instruments called *rubrics* (i.e., performance measures). Each rubric measures a different facet of teaching. Teachers are ranked differently depending on the criteria used for different rubrics (Pittman, 1992); they might be high on one dimension and low on another. Different instruments are needed to measure the multifaceted process of teaching. One instrument is like a single frame of a motion picture, capturing only some of the action (Burry, in press). Each frame is an important part of the overall picture and yet one frame captures only one aspect.

The motion picture analogy can also be used to illustrate the use of the model. Often we need only one picture to illustrate a facet of teaching. For example, a classroom observation instrument provides us with usable information about what goes on in the classroom. It provides us with only classroom information about a teacher(s) and students during a specified period of time. It does *not* provide us with information about what students have learned, how the teacher perceives his/her own teaching, how the teacher evaluates student learning and provides feedback to students, the ability of the teacher to reflect on his/her teaching practices and professional habits, or how the teacher assembles content and materials to facilitate student learning of science process and content skills. Each teaching facet depicts a different aspect of teaching. By assembling these facets into one unit, we hope to provide a more comprehensive picture of teaching, akin to a motion picture. As explained throughout this manuscript, the theoretical base combines the current constructivist perspective that pervades the science education community, the findings of the National Center for Research on Science Teaching, Scriven's "Duty Based Evaluation Model," the work on novice through expert teaching, as well as the project literature reviews. At the time of this writing, there are no constructivist classroom observation instruments and there is no expert science teaching, or teaching, evaluation model that groups evaluation instruments together to provide a more comprehensive picture of what expert teaching looks like.

Another unique facet is that the ESTEEM also includes student outcomes that are conceptually based indicators of student understanding as a facet of describing expert science teaching. Leaders in the field of science education and effective teaching, including Senta Raizen (National Center for the Improvement of Science Education), Carolyn Evertson (Peabody College/Vanderbilt University), David Berliner (Arizona State University), Robert Yager (University of Iowa), and Joe Novak (Cornell University), provided excellent consultation on specific aspects at various times for the Expert Science Teaching Project.

The ESTEEM is theoretically and empirically based. However, it is highly unlikely that a teacher would be considered an expert on all of the rubrics, especially as the majority of the expert science teachers nominated in this study, exemplars from seven states, were not constructivist science teachers (see discussion later in the manuscript).

It is our intention that the instruments be used after thorough training as an impetus for professional development. They may be self-administered, peer-administered, or administered by an outside evaluator. The instruments should be used one or two at a time beginning with the *Classroom Observation Rubric* and the *Student Outcomes Assessment Rubric*. Over the past three years the ESTEEM staff has developed five instruments and has plans for at least two more. The first two rubrics designed in 1991 and 1992 are the *Classroom Observation Rubric* and the *Student Outcomes Assessment Rubric*. These two rubrics are used together to evaluate a single science lesson. Three additional rubrics designed in 1993 were piloted with the Alabama Science Teaching and Learning project, funded by the Dwight D. Eisenhower Mathematics and Science Program (PL 100-297) overseen by the Alabama Commission on Higher Education. The *Concept Mapping Rubric* evaluates formative and summative student conceptual understanding. The *Science Grading Practice Assessment Inventory* is a self-report instrument designed to assess the degree to which a teacher feels competent with grading various aspects of student achievement. The third instrument is the *Teaching Practice Inventory* which is a self-report measure of the individual behaviors listed on the *Classroom Observation Rubric*. This instrument is a nice way to determine the awareness level of science teachers relative to the constructivist/expert science teaching perspective. The two new rubrics, *Reflective Teaching Practices* and *Instructional Design*, are in draft form and will be piloted during the 1993-1994 school year.

We patterned the ESTEEM rubrics after the analytical scoring system developed by Spandel and Stiggins (1990) to assess writing skills. The rubrics utilize a behaviorally anchored rating scale to assess the teaching practices at five-point intervals. The descriptions of teaching practices are described at points "5," " 3," and "1." Points "4" and "2" are also usable, but they are not usually operationally defined; they are interpreted as points between the definitions for points "5," "3," and "1." All teaching practice descriptions on the rubrics have the same rating format, to facilitate the development of an ESTEEM profile.

The ESTEEM rubrics were written to describe the ideal practices of expert science teachers from a constructivist and expert teaching perspective. A rating of "5" would indicate an expert level. Likewise, a "3" would indicate a capable, experienced teacher and a "1" would indicate poor constructivist teaching practices. A "0" can be used if no information is provided. The "5" rating is intended to be used as a quantification of the expert level as defined by Leinhard

and Greeno (1986) and Berliner (1987). This means that is it not easily obtained. It is highly doubtful that a teacher would receive all "5s" on any one rubric let alone a "5" average for all rubrics.

Classroom Observation Rubric: The *Classroom Observation Rubric* was developed using data collected with the comprehensive written scripts from science classes taught by nominated expert science teachers from seven states. It was developed by a panel composed of experienced science educators and researchers Burry, Bolland, Sunal, Pittman, Sunal, Turner, Rice, Hedgepath, and Zhang (1991). Data came from transcripts and interviews of 46 fourth- through eighth-grade teachers nominated as expert science teachers. These data, along with an extensive literature review, assisted the panel in identifying behaviors documentable on the *Science Classroom Observation Rubric.*

The panel placed observable practices in categories, representing primarily a constructivist and expert teaching perspective on classroom instruction, which became the *Classroom Observation Rubric.* The *Classroom Observation Rubric* was then used to evaluate the transcripts of the 46 nominated expert science teachers.

For example, under Category 2, Content-Specific Pedagogy, the panel assigned a "5" to the teaching practice "Teacher is constantly making the content of the lesson relevant to student understanding." A "3" was linked with the teaching practice "Teacher sometimes makes the content of the lesson relevant to student understanding." "Teacher does not make the content of the lesson relevant to student understanding" received a rating of "1." The numbers "2" and "4" may also be used, even though they are not defined. If the practice is not performed, the rating is a "0". Under each teaching practice description, the panel provided examples of how that particular behavior might be exhibited in the classroom. Table 2 is an excerpt from Category 2.

Table 2

Excerpt from the *Classroom Observation Rubric*

Category 2: Content Specific Pedagogy

5 Teacher is constantly making the content of the lesson relevant to student understanding.

F. The lesson mainly focuses on activities that relate to student understanding of concepts. Student relevance is always a focus.

G. Students have an opportunity to experience the relationship of the concept to their everyday lives.

H. During the lesson the teacher appropriately varies methods to facilitate student conceptual understanding; i.e., discussion, questions, brainstorming, experiments, log reports, student presentations, lecture, demonstration, etc.

I. Teacher consistently moves students through different cognitive levels to reach higher order thinking skills.

J. Content and process skills are integrated.

K. Concepts are connected to the evidence.

The transcripts from the nominated 46 expert science teachers were analyzed using *The Classroom Observation Rubric*. Table 3 illustrates the means, standard deviations, and the reliability coefficients.

Table 3

Means, Standard Deviations, and Reliabilities for Science Classroom Observation Rubric (SCOR) Total and Factors

Variable	$\bar{X}$	SD	Reliability
SCOR Total	57.30	12.69	.91
Facilitating the Learning Process (F1) (5 behaviors)	15.33	4.81	.84
Content Specific Pedagogy (F2) (6 Behaviors)	19.15	5.12	.89
Contextual Knowledge (F3) (3 Behaviors)	9.28	2.70	.87
Content Knowledge (F4) (4 Behaviors)	13.54	3.72	.80

The mean is 57.30, the maximum score is 90, the median is 61, the 25th percentile is 50, the 75th percentile is 66, and the standard deviation is 12.69. Even the smallest (three-item) factor has a reliability of .80. Highest factor reliability is .89. Overall reliability is .91. Note, the mean of 57.30 for the 46 nominated expert science teachers suggests that most of these teachers were not constructivists. Even though these teachers were nominated by college and university

faculty, state department of education personnel, and regional office personnel, it appears that the constructivist perspective is not widely held by either the nominators or their nominees (our subjects).

Groups that were artificially formed using the top and bottom quartiles, and *t*-tests were computed and found statistically significant at the .01 level with Bonferroni correction for the alpha level for all four categories and the observation rubric total (Pittman, 1992). This means that the top and bottom quartiles were significantly different from each other for each category and for the total. Although all participants were nominated as expert science teachers, nevertheless the top quartile and the bottom quartile operated significantly differently from each other. This information contributes to the construct validity of the Classroom Observation Rubric.

Table 4 illustrates the results from the assessment of the 46 nominated expert science teachers, which were factor analyzed using a principle component solution with an orthogonal rotation of four factors. The final factor solution accounted for 71.3 percent of the variability with the following four factors, subscales, which are labeled categories: (1) Facilitating the Learning Process; (2) Content-Specific Pedagogy; (3) Contextual Knowledge; and (4) Content Knowledge. All of the 18 teaching practices have factor loadings from .8538 to .5596. Factor loadings are interpreted in the same manner as correlation coefficients, are above .7000.

Table 4

Principle Component Solution with a Varimax Rotation for the Classroom Observation Rubric

	Factor 1	Factor 2	Factor 3	Factor 4	Final Commonality Estimates
1	.8433				.81
2	.8732				.83
3	.7034				.76
4	.6216				.70
5	.7017				.51
6		.7433			.73
7		.4712			.70
8		.8096			.79
9		.6567			.61
10		.5896			.84
11		.8031			.62
12			.5597		.55
13			.8538		.74
14			.7030		.77
15				.8144	.69
16				.8309	.78
17				.7765	.69
18				.8449	.76
Sum of the square factor loadings	3.5818	3.7451	2.2573	3.2497	12.84
% Variance	19.9	20.8	12.54	18.1	71.34

NOTES: **Values less than .5896 have been eliminated**

Results of factor analysis of Classroom Observation Rubric using principal components solution with Varimax factor rotation, N = 46.

The Student Outcome Assessment Rubric: Two open-ended student questions comprise the second rubric, (Burry & Bolland, 1991). The *Student Outcome Assessment Rubric* is a standardized performance measure with instructional content validity, because the students are evaluated on information taught only in their class. The roots of the *Student Outcome Assessment Rubric* are in the constructivist concept of "teaching for meaning," which suggests that the rubric

also has construct validity. The first open-ended question was adapted from material written by Angelo and Cross (1991). The evaluator (the person who observes the lesson and administers the student questions) evaluates the students' responses, which are then compared to the criteria on the rubric and designated as a "1," " 2," "3," "4," or "5". A "0" may be used if there is no response. Unlike the previous rubric level, "2" is defined due to the nature of the content being assessed.

Table 5 is an excerpt from the *Student Outcome Assessment Rubric*. Here we code whether the student captured the main idea as it was presented during the lesson.

Table 5
Excerpt from the *Student Outcome Assessment Rubric*

Capturing the Main Idea
5. The response states the main idea and provides details, descriptions, or explanations that indicate the student did not just copy or regurgitate the main idea. The response indicates the student understood the "big picture" surrounding the main idea. Response may go beyond the idea as discussed in class.
3. The response states the main idea, with no elaboration. The statement may appear to be book-related.
2. The response was related to the main idea but did not indicate that the student understood that there was a main point. The student stated facts (discussed in the lesson) about the main point without describing the big picture. For example, give a "1" rating to "Cirrus is a type of cloud," if the lesson was about the three types of clouds and their shapes.
1. The student's response has little or no relationship to the main point of the lesson. The response is about a different topic or an aspect of the broader topic. For example, "Humans have two arms" should be rated "1" if the lesson was about the endocrine system.

An interrater agreement index (six judges) of .78 was calculated on a randomly selected behavior. The index is used to evaluate performance data across raters in an algebraically equivalent set of formulas (Burry, Shaw, Chissom, & Laurie, in press). An interrater index of 0 indicates no agreement among judges, and 1.00 is perfect agreement. An index of .78 is very respectable for data on a performance measure.

Interrater agreement indices were calculated on random samples of the data for the *Student Outcome Assessment Rubric*. The interrater agreement indices are .88 for the main idea question and .79 for the second question, the relevance question. Both of these indices are more than acceptable. At the time of the first writing the data were designed as ordinal, so measures of internal consistency were not calculated. Student data were aggregated, and medians were calculated for each teacher for use in the data analysis reported in the next section.

High and low groups were created for statistical analyses using teachers whose student proportions were in the top and bottom quartiles (Pittman, 1992). As suggested by Hinkle, Wiersma, and Jurs (1988), the group means for top and bottom quartile proportions were used to calculate a *Z*-test for proportions. The main idea *Z* value is 7.50, and the relevant question *Z*-value is 7.17. The value of 2.575 is the alpha level needed for these tests to be statistically significant at the $p < .01$ level. The *Student Outcome Assessment Rubric* thus discriminates well between the proportions of students in top and bottom quartiles of nominated expert science teachers. The rubric demonstrates both content and construct validity.

Teaching Practices Inventory: This inventory assesses each of the behaviors listed from the practices in the *Classroom Observation Rubric.* This inventory was adapted for use in the ESTEEM by Turner and Burry (1992). Teachers are asked to rate themselves on a frequency scale from "1" to "5" of how often they perceive they use each behavior (See Table 6). The 30 behaviors on the *Teaching Practices Inventory* have a Cronbach alpha reliability coefficient of .93 and a four factor principle factor solution with a varimax rotation accounting for 61% of the variability. The factor solution is very similar, but not identical to the *Classroom Observation Rubric* as would be expected, because two different perceptions are evaluated (Burry & Shaw, 1988).

Table 6

Excerpt from the *Teaching Practice Inventory*

Using the following response scale rate the frequency with which you feel that you use the following science teaching practices:

NEVER (1)
OCCASIONALLY (2)
SOME OF THE TIME (3)
MOST OF THE TIME (4)
ALMOST ALWAYS (5)

1. Your students are responsible for their learning.
2. Your students are actively engaged in initiating examples throughout the lesson.
3. Your students are actively engaged in asking questions throughout the lesson.
4. Your students are actively engaged in suggesting activities throughout the lesson.
5. Your students are actively engaged in implementing activities throughout the lesson.

Science Grading Practice Inventory: The *Science Grading Practice Inventory* was developed to assess the degree to which teachers felt skilled in using student evaluation procedures appropriate for science. The forerunner of the *Science Grading Practice Assessment Inventory* was the *Grading Practice Assessment Inventory* written by Malcolm and Burry (1990). The inventory was adapted for the science classroom by Burry (1993). The teacher responds using a scale of "0" "not at all skilled" to "9" "highly skilled" to 66 items. Table 7 illustrates a set of items from the inventory.

Table 7

Excerpt from the *Science Grading Practice Assessment Inventory*

SCIENCE GRADING ASSESSMENT INVENTORY

NOT AT ALL SKILLED 0 / 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 **HIGHLY SKILLED**

41. Using science notebooks.
42. Using laboratory worksheets.
43. Using individual hands on activities.
44. Using group hands on activities.
45. Using individual class presentations.
46. Using group class presentations.
47. Using the end of chapter questions.
48. Using informal teacher observations.

The Cronbach alpha reliability coefficient is .98. A principle component factor analysis with an oblique rotation produced six factors that accounted for 85% of the variability.

Concept Mapping Rubric: This is the newest of the rubrics and it is still in the embryonic stage. It was developed using much of Novak (1990) and Novak and Gowin's (1984) original work in the area of concept mapping. The rubric was piloted with eight teachers from the Alabama Science Teaching and Learning Professional Development Project funded by the Dwight D. Eisenhower Mathematics and Science Education Program. A committee of science educators, cognitive psychologists, educational psychologists, and educational researchers worked on this first draft (Burry et al., 1993). Based upon the results of a study done with these teachers and their students, a simplified version is currently in draft form (Bolen, Lacefield, & Burry-Stock, 1993). An excerpt describing the "5" level rating appears in Table 8. It is assumed that students are taught the procedures for doing a concept map as were the students in the study (Burry-Stock, 1993).

Table 8

Concept Mapping Rubric for a Rating of "5"

Category 1: Concepts

5	a.	Approximately 90 percent of the words/concepts from the key map are present. The number of key words is divided by the total number of key words on the word list.
	b.	Fifty percent or more of the words/concepts are other than those from the key map.* The number of non-key words is divided by the total number of words on the concept map.

* Those concepts and relationships not included on the criterion list. These concepts are called non-key concepts.

Category 2: Simple Concept Relations:

5	c.	Approximately 90 percent of the relationships between two key words are indicated by a connecting line.
	d.	Connecting lines are labeled with a word or symbol.
	e.	The relationships between the key concepts are meaningful.
	f.	The relationships between non-key and key concepts or between two non-key concepts are indicated by a connecting line.*
	g.	The relationships between non-key and key concepts or between two non-key concepts are labeled with a word or a symbol.*
	h.	The relationships between non-key and key concepts or between two non-key concepts are meaningful.*

Category 3: Conceptual Relations

5	i.	The map shows a meaningful pattern. Each key concept that is more specific and less general than other key concepts is drawn/written to demonstrate the relationship.
	j.	Each non-key concept is shown in its appropriate place in the pattern.*

Category 4. Cross Links

5	k.	The map shows significant and meaningful connections between one segment and another segment.
	l.	The map shows significant meaningful connections between one key segment and another non-key segment or between two non-key segments.*

Category 5: Conceptual Understanding

5	m.	The concepts are arranged to show deep understanding.

Correlational and other Research Studies

During the first year of the CREATE Expert Science Teaching Project 46 teachers were selected from a list of over 150 nominated expert science teachers. Nominators, including school superintendents, teacher educators, state departments of education, and intermediate education agencies were identified and information concerning our project was distributed to them. Nominators submitted names of teachers for participation in the study. The nominations were reviewed and the site team work was begun with teachers from Alabama, Florida, Georgia, Illinois, Louisiana, North Carolina, and Tennessee. Participants were selected after correspondence and an initial telephone interview with the teacher and the teacher's principal.

Table 9 illustrates the state representation of our teachers, Table 10 illustrates the number of years of experience (note that beginning teachers were also included so as to cover the experiential continuum), and Table 11 illustrates the teachers' level of education. Note that Table 10 includes exemplary first year teachers as this study was also concerned with the developmental aspects of teachers.

Table 9

Sample Characteristics of Nominated Expert Science Teachers

State	Frequency	Percent*
Alabama	19	4.13
Florida	4	8.7
Georgia	10	21.7
North Carolina	1	2.2
Tennessee	3	6.5
Louisiana	4	8.7
Illinois	5	10.9

* Totals more than 100 due to rounding error.

Table 10

Years Teaching

Alternative	Frequency		Percent	Cumulative Frequency	Cumulative Frequency
1-2	3	6.5	3	6.5	
3-4	3	6.5	6	13.0	
5-6	2	4.3	8	17.4	
7-8	6	13.0	14	30.4	
9-10	2	4.3	16	34.8	
11-13	8	17.4	24	52.2	
14-16	6	13.0	30	65.2	
17-20	3	6.5	33	71.7	
21-30	11	23.9	44	95.7	
31 or more		2	4.3	46	100.0

Table 11

Highest Degree

Alternative	Frequency	Percent	Cumulative Frequency	Cumulative Frequency
Bachelors	5	10.9	5	10.9
Bachelors+	9	19.6	14	30.4
Master	13	28.3	27	58.7
Master+	17	37.0	44	95.7
Specialist	2	4.3	46	100.0

Site Team Members

Nine site team members, consisting of science teachers, science education faculty, and research faculty were trained by the project director at The University of Alabama to conduct the interviews, do the classroom observations, evaluate student outcomes, and evaluate the classroom observation. All members have received about 20 hours of training to participate in this study.

Demographic Variable Analyses

Demographic variables, time spent on various activities, and cognitive levels were the first variables of interest. Demographic variables were also used in analyses with the *Classroom Observation Rubric* and the *Student Assessment Outcome Rubric*. Each of the analyses will be presented followed by a summary table.

The amount of time spent on all of the activities was obtained from the transcripts of the classroom observations. The distribution of the time spent on various activities was irregular, because some teachers spent almost the entire class time on one or two types of activities, and very little or no time on others. The time spent on these activities varied from teacher to teacher. Consequently, the distribution of the times was not linear, and the data were smoothed using the natural log transformation. All of the data analyses using the activity code times were transformed. The activity code and cognitive codes are listed in Figure 2.

Figure 2. Classroom Activity Code Categories

Description
Content Development: Teacher Presentation of Content
Content Development with demonstration: Teacher Presentation with Content Development
Review (may include informal assessment)
Interaction
Group Work Including Pairs with Interaction
Group Work with Hands-on Activities
Small Group Instruction
Group Presentation
Directions for Assignments
Individual Seatwork
Student Presentations
Evaluation (Examination, tests, quizzes)
Managing Behavior
Looking over Student Work
Administrative Routines
Non-academic Activity
Transitions
Disciplinary Incidents
Waiting Time

Contingency coefficients were calculated for all of the activity codes and the demographic data. Contingency coefficients were used because the data were a combination of categorical and interval levels. The data were also ipsative, in that individuals were comparted with a criteria not a norm. Because each teacher taught a lesson that was unique with different activities, teachers did not use the same or all of the activity codes. The activity code sample cells were different for each teacher. This means that the time spent on various activities was specific to each teacher's lesson. Therefore, some of the activity sample cells had a small sample size. The calculation of contingency coefficient from chi-squares provides a means of comparing the data. The contingency coefficient is calculated using the square root of the chi-square value divided by the sample size. The data were then comparable because the sample size is taken into consideration

in the calculations. Consequently, chi-squares and contingency coefficients were selected for this level of analysis. The use of contingency coefficients allows for interpretations that are similar to correlation coefficients.

All possible combinations of activity codes and demographic variables were calculated. Only those contingency coefficients appearing in the top 15% of the frequency distribution of all possible combinations of these data were used. These contingency coefficients are considered to be salient. The salient coefficient of interest is equal to or greater than .85. Table 12 illustrates the salient contingency coefficients for the demographic variables with the natural log transformations of the time spent with the activities and on cognitive levels.

Table 12

Contingency Coefficient Analysis: Years of Teaching Experience with Natural Log Transformations of the Time Spent on Activities and At Cognitive Levels ≥ (c .85)

Demographic Variable	Activity/Cognitive Level	Contingency Coefficient
Years Teaching	Content Development	.93
	Content Dev. with Demo.	.88
	Discipline	.87
	Examinations	.85
	Group Hands On	.91
	Interaction	.92
	Looking over Student Work	.92
	Management	.89
	Seatwork	.86
	Student Hands On	.90
	Transition	.89
	Recall	.93
	Comparison	.85
	Influence	.85
	Evaluation	.87

Seven of the 17 demographic variables also revealed salient contingency coefficients. The variables were: (1) hours of inservice; (2) grade level; (3) student ethnic configuration; (4) social economic level; (5) number of students in a class; (6) student ability level; and (7) district type.

Because we are interested in the contexts of our 46 expert teachers, further analyses were done with the demographic variables (Burry & Oxford, 1993). Analyses of variance were done on the *Classroom Observation Rubric* with demographic variables as illustrated in Table 13.

Table 13

F tests for *Classroom Observation Rubric* with Demographic Variables

Demographic Variable	F	Prob	Description of Means
Teacher's Ethnic Background	9.31	.00**	Higher for white
Years of Teaching Experience	.97	.48	Increases with Experience (Drops for 11-16 years)
Years in Current Setting	1.10	.38	Irregular pattern
Hours of Science In-Service	.60	.73	Slightly higher for more hours
Student Ethnic Configuration	4.15	.01**	White, White and Black, Asian Highest
Social Economic Level	.24	.94	High and Low to High highest
Number of Students	.24	.94	Highest for Low to high in class
Student Ability Level	.56	.69	Low to High one class highest
School Population	.22	.88	Highest with Largest Schools
District Type	1.70	.16	Suburban highest
District Population	1.94	.16	Highest with Large Districts
Grade Level	.82	.54	Lower grades are higher
Length of Lesson	4.02	.00**	Highest at 41-50 minutes
Teaching Certificate	7.62	.00**	Higher for elementary
Highest Degree	.41	.80	Increases with degrees
Number of Subjects	2.73	.06	Three subjects
Teaching Model	.18	.67	No differences

** $p > .01$ ** $p > .05$

Demographic variables that appear significant are: (1) years of teaching experience, which also appears across 15 activity codes and cognitive levels; (2) number of hours of science in-service; (3) number of students in the class, surprisingly favoring larger classes (30 or more); (4) district population, favoring the larger districts; (5) length of science lessons, favoring 41-50 minutes; (6) student ethnicity; (7) grade level; (8) degree status; and (9) current teaching setting. As our results confirm, it is reasonable to expect that years of experience would contribute, as specialists (Berliner, 1987; Carter et al., 1988; Leinhardt & Greeno, 1986) suggest that experience determines expertise. Our results confirm the expectation that the better teachers have more science in-service training. The number of students in the class, district size, and district type appear to be related to where the better teachers are employed. As we might have anticipated, the better teachers teach in suburbs with the largest district populations. This study reveals that better teachers are in classrooms where there are more white students.

Elementary school teachers have higher means than secondary school teachers, and the means go down consistently as the grade levels go up. This indicates that elementary school practices are more aligned to the constructivist perspective than are secondary school practices. The higher the degree status, the better the teacher, according to this study. In addition, the longer the teacher is in the current setting, the better the teacher turns out to be.

The *Classroom Observation Rubric* total and category scores as well as the *Student Outcomes Assessment Rubric* were correlated with the log transformation of the activity code times as illustrated in Table 14.

Table 14

Summary of Activities and Cognitive Correlations Appearing on SCOR and SOAR $r \geq .40$

Activities/Cog. Level	r Class Obs	r Stud Out
Group Work	.45 Total	.85 Total
Content Dev with Demo	-.62 Cat 1	.65 Total
Content Dev with Demo	-.62 Cat 1	.51 Main Idea
Group Work	.52 Cat 1	.85 Int/Rel
Review Recitation	.62 Cat 2	.65 Main Idea
Group Work	.50 Cat 2	.85 Int/Rel
Seat Work	.70 Cat 3	-.51 Int/Rel
Seat Work	.45 Cat 4	-.51 Int/Rel

The salient correlations that appear across both the *Classroom Observation Rubric* and the *Student Outcome Assessment Rubric* are referred to in Table 14. Group work is important to teaching and student behavior. Group work is also highly correlated with the interest/relevancy question. This suggests that group work is facilitating the conceptual understanding that constructivists are promoting. Category 1 of the observation instrument is facilitating the learning process. This relationship also suggests that the *Classroom Observation Rubric* and *Student Outcomes Assessment Rubric* have this point in agreement. The natural log transformation of the time spent with content demonstration suggests a negative correlation with the classroom observation Category 1, because this category is promoting conceptual understanding where the student, not the teacher, is to take the responsibility for the learning experience. Most of our expert nominated teachers did not demonstrate this variable well. However, the content development did foster conceptual understanding as measured by the total *Student Outcome Assessment Rubric* summation and the individual item, main idea. Review and recitation were highly correlated with the classroom observation Category 2 Content Specific Pedagogy and the Main Idea. This means that the focus on conceptual understanding, structuring relationships with concepts to students' everyday lives, the teacher's conceptual monitoring, and integrated science and process skills are important to understanding the main idea of the lesson. Group work was also correlated with this Category. This suggests that the content specific pedagogy described in the *Science Classroom Observation Rubric* facilitates both review and recitation, recall, and

If we want students to learn conceptually, we are going to have to teach in a way that facilitates a higher level of student understanding. There is more to learning than information that is memorized at a recall level, as this type of information may be good for classroom learning, but does not provide a link to making science relevant and useful. In today's society, more than ever, it is critical that students be scientifically literate.

The ESTEEM can be implemented with teachers on a self-report, peer, or outside evaluator professional development guide. We intended the model to be used as a multi-faceted approach for improving science teaching. Both teacher and student behaviors are identified on the five existing rubrics. All of the rubrics demonstrate excellent reliability and validity properties, which strongly suggest that the instruments can be used for their intended purpose to measure expert science teaching from a constructivist perspective for professional development. Teachers should employ one or two of the rubrics at a time starting with the *Classroom Observation Rubric* and the *Student Outcomes Rubric*. It may take several years for a teacher to feel that he/she has mastered the practices described on the *Classroom Observation Rubric* as is illustrated by the research in this manuscript. Many, if not most, science teachers do not teach from a constructivist perspective. The research from this study indicates that there may be good science teaching and there may be many teachers who are experts in their content, but most of the teaching is done from a teacher centered approach and much of the content that is disseminated is at the recall or lower end of the high-order-thinking skills taxonomy. This is suggested by the relatively low percentage of students scoring at the upper cognitive levels on the *Student Outcomes Assessment Rubric*.

Our efforts to develop instruments and define characteristics seem to echo the science community's interest (Loucks-Horsley, et al., 1990; Yager, 1991). The model that we have developed is a way of getting teachers to work with each other for professional development that will enhance teaching and learning. The ESTEEM reflects theoretical premises from Scriven (1991) and Guba and Lincoln (1989). Given that many of our expert science teachers do not teach from a constructivist perspective, it may be that one of the best ways to encourage this practice is to train teachers and have them work together to strive for excellence in the classroom.

It is critical that we involve teachers in the development, dissemination, and implementation of the ESTEEM. For as Scriven (1990) said "In research literature, the definition of success in teaching often appears to be circular, i.e., that any particular description of success emanates from what the researchers themselves have decided, and not from what behaviors actually cause student learning" (p.20). Not only do we want to verify through research what needs to happen, but we also need to validate that it has happened and that it works or it does not work. We are attempting to do this. If we are going to enhance science for all as suggested by the American Association for the Advancement of Science in *Science for All Americans: A Project 1061 Report on Literacy Goals in Science, Mathematics, and Technology* and *America 2000* (1991), we need revised and new methods of creating school environments to enhance teaching for learning.

If we in America are to provide an equal opportunity for all students to be scientifically literate as suggested by the authors of *Project 2061* and *America 2000*, we need to find ways to activate the suggested outcomes of the noted researchers cited in this paper. If we are to be successful in preparing students for the 21st century we must change our science teaching practices. This does not mean that all of the old methods are of no use. What it does mean is that there are ways to improve teaching to make science more meaningful and useful to students

conceptual understanding. Seat work had a high correlation with the classroom observation Category 3 Contextual Knowledge. This is not surprising, as the variables subsumed in this category deal with confronting misconceptions, demonstrating good interpersonal relations with students, and using awareness skills to modify lessons. However, the negative correlation of seat work with the *Student Outcomes Assessment Rubric* interest/relevancy suggests that seat work is directed more towards the learning of facts than concepts. The same type of relationship exists between seat work and Category 4 of the classroom observation "Content Knowledge." This category deals with the accuracy of the lesson, the in-depthness and comprehensiveness, the integration of concepts, generalizations, and skills in a coherent fashion, and the unique nature of exemplars and metaphors. This category is related more toward the lower cognitive levels. Again, seat work is negatively correlated with the SOAR interest/relationship item, suggesting that seat work is geared toward the recall level only.

Other important salient correlations of the *Student Outcome Assessment Rubric* with learning activities appear in Table 14. The *Student Outcomes Assessment Rubric* total correlates with content development accompanied by demonstrations. This suggests that demonstrations do foster conceptual understanding. It also correlates with group work. This is an interesting point, as group work includes the interaction of students with each other, as well as students with the teacher. It also suggests that students working with students help each other to learn. Student understanding and conceptualization can be facilitated by group activity, as suggested by Anderson and Roth (1989), Loucks-Horsley et al, (1990), and Yager (1991). The main idea is associated with content development inversely related to review and recitation, and positively related to evaluation. Interest/relevancy was positively related to group work, student hands-on and non-academic time. Non-academic time was coded by the site team people when it was apparent instruction had not occurred. However, it appears that the observer does not always know when learning transpires. Inverse relationships of interest are management, review/recitation, examination, and seatwork.

Conclusions and Next Step

The data reflect students' and teachers' behaviors from classrooms of 46 expert science teachers. Instruments used proved reliable and valid. One major limitation of the sample is that we do not know on what basis the teachers were considered to be expert. We have learned from this study that expert teachers are defined by the criteria on which they are evaluated. There is only about a 50% agreement in the top quartile of the teachers sorted by the *Classroom Observation Rubric* and sorted by the *Student Outcome Assessment Rubric* (Pittman, 1992). This alone should alert us to being very cautious about what is considered to be expert.

The theoretical base for our work is provided by the constructivist movement, Scriven's Duty Based Evaluation, and the novice through expert literature. The constructivist movement is very strong in science (Yager, 1991), but is also prevalent in other subject areas and facets of education (Evertson & Murphy, in press; Guba and Lincoln, 1989). The mean for our 46 teachers on the *Classroom Observation Rubric* is 57 out of 90. This suggests that our nominated expert science teachers are not well informed constructivists. The proportion of students scoring on the upper levels of the *Student Outcome Assessment Rubric* are not very high, which also suggests that our nominated expert science teachers are not teaching at a particularly high conceptual level.

in their everyday lives. The CREATE Expert Science Teaching Evaluation is an attempt to implement a constructivist/expert teaching approach to teaching and learning that we hope will create an impact on student learning that enables children to be better able to understand and cope with the scientific phenomena that affect our daily lives. We hope to assist the nation in finding ways that will produce scientifically literate citizens.

Note

[1] Excerpts from this manuscript appear in the *ESTEEM Manual* and monograph entitled *Measuring Excellence in Science Teaching: Expert Science Teaching Evaluation (ESTEEM)*. All of the ESTEEM instruments are included in the *ESTEEM Manual* and are copyrighted by Judith A. Burry-Stock. The manual, the monograph, and an extensive literature compilation may be purchased by writing CREATE at the Center for Evaluation, Western Michigan University, Kalamazoo, MI 49008-5178 or calling 615-387-5835.

References

America 2000: An education strategy sourcebook. (1991). Washington, DC: U.S. Government Printing Office.

American Association for the Advancement of Science. (1989). *Science for All Americans: A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, DC: Author.

Angelo T.A., & Cross, K.P. (1993). *Classroom assessment techniques.* San Francisco, CA: Jossey-Bass Publishers

Anderson, C. W., & Roth, J. J. (1989). Teaching for meaningful and self-regulated learning of science. In J. Brophy (ed.), *Advances in research on teaching: Vol. 2. A research annual.* 1, 265-309. Greenwich, CT: JAI Press, Inc.

Applegate, J. L., & Sypher, H. E. (1988). A constructivist theory of communication and culture. In Y. Y. Kim & W. B. Gudykunst (eds.), *Theories of intercultural communication* (pp. 41-65). Newbury Park, CA: Sage.

Ball, D. L. (1988). *Prospective teachers' understanding of mathematics: What do they bring with them to teacher education*? Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Barth, R. S. (1988). School: A community of leaders. In A. Lieberman (Ed.), *Building a professional culture in schools.* New York: Teachers College Press.

Beck, L. G., & Murphy, J. (in press). *Understanding the principalship: A metaphorical analysis from 1920 to 1990.* New York: Teachers College Press.

Berliner, D. C. (1987). In pursuit of expert pedagogue. *Educational Researcher, 15*(7), 5-13.

Bliss, J.R., Firestone, W.A. & Richards, C.E. (Eds.) (1991). *Rethinking effective schools: Research and practice.* (Rutgers Symposium on Education). Englewood Cliffs, JF: Prentice -Hall.

Bolan, L., Lacefield, J., & Burry-Stock, J. (1993). *Expert science teaching educational evaluation model (ESTEEM); Concept Mapping Rubric: Measuring Student Conceptual Understanding.* Paper presented at the Mid-South Educational Research Association MSERA, New Orleans, LA.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal, 26(4)*, 473-498.

Brennan, E.B., Zhang, Z., Slater, W., & Bolland, K. (1993) *Literature Review for the Expert Science Teaching Evaluation (ESTEM)* J. Burry (Ed.), Kalamazoo, MI: Center for Evaluation.

Burry, J. A. (in press). Perspectives on methods and procedures for assessing various stages of teacher effectiveness. *Journal of Personnel Evaluation in Education.*

Burry, J.A. (1993). *Science grading practice assessment inventory (SGPAI).* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry-Stock, J.A. (1993). *Expert science teaching educational evaluation model (ESTEEM) Manual.* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry, J. A. & Bolland, K.A. (1991). *Student outcomes assessment rubric (SOAR).* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry, J. A., Bolland, K. A., Pittman, K., & Brennan, E. (1992). The process of developing a composite of expert science teaching. In D. C. Berliner (Chair), *A search for excellence: What does expert science teaching really look like?* Symposium conducted at the Annual Meeting of the American Educational Research Association, San Francisco.

Burry, J. A., Bolland, K. A., Sunal, D., Rice, J., Turner, M., Hedgepath, D. & Zang (1991). *Classroom observation rubric.* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry, J.A., & Oxford, R. L.(1993). *Measuring Excellence in Science Teaching: the Expert Science Teaching Evaluation (ESTEM).* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry, J.A., Roskos-Ewoldsen, B., Iran-Nejad, A., Oxford, R., Bolen, L. (1993). *Concept Mapping Rubric.* Kalamazoo: MI, The Evaluation Center, Western Michigan University.

Burry, J., & Shaw, D. (1988). Teachers and administrators differ in assessing teacher effectiveness. *Journal of Personnel Evaluation in Education, 2,* 33-41.

Burry, J., Shaw, D.G., Chissom, B.S., & Laurie, C. (in press). Rater agreement indices for performance assessment. *Educational and Psychological Measurement.*

Carter, K., Cushing, K., Sabers, D., Stein, P., & Berliner, D. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39*(3), 25-31.

Conway, J. A., & Jacobson, S. L. (1990). An epilogue: Where is educational leadership going? In S. L. Jacobson & J. A. Conway (eds.), *Educational leadership in an age of reform.* New York: Longman.

Crockett, W. (1965). Cognitive complexity and impression formation. In B. A. Maher (ed.)., *Progress in experimental personality research* (pp. 47-90). New York: Academic Press.

David, J. (1989). *Restructuring in progress: Lessons from pioneering districts.* Washington, DC: National Governors' Association.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind over machine*. New York, NY.: Free Press a division of Macmillan In.

Dwyer, C. A. (1991). Measurement and research issues in teacher assessment. *Educational Psychologist, 26*(1), 3-22.

Elmore, R. F. (1989). *Models of restructured schools*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Evertson, C. M. (1990). NSTA's Scope, Sequence & Coordination Project: Critical questions for a research agenda. In J. A. Shymansky and W. C. Kyle, Jr., *Establishing a research agenda: The critical issues of science curriculum reform*. Report of a conference sponsored by the National Association for Research in Science Teaching.

Evertson, C. M., & Murphy, J. (in press). Beginning with the classroom: Implications for redesigning schools. In H. Marshall (Ed.), *Supporting student learning: Roots of educational change*. Norwood, NJ: Ablex.

Evertson, C. M., & Randolph, C. H. (in press). Teaching practices and class size: A new look at an old issue. *Peabody Journal of Education*.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (1988). *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin Company.

Hollingsworth, S. (1990). *Teacher educator as researcher: An epistemological analysis of learning to teacher reading*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Hopkins, K.D., Stanley, J. C. & Hopkins, B.R. (1990). *Educational and psychological measurement and evaluation*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Kelly, G. (1955). *The psychology of personal constructs*. New York: North.

Kyle, W. C. (1991). The reform agenda and science education: Hegemonic control vs. counterhegemony. *Science Education, 74*, 403-411.

Lambert, M., & Clark, C. M. (1990). Expert knowledge and expert thinking in teaching: A response to Floden and Klinzing. *Educational Researcher, 19*(4), 75-95.

Leinhardt, G. (1990). Capturing craft knowledge in teaching. *Educational Researcher, 19*(2).

Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology, 78*(2), 75-95.

Loucks-Horsley, S., Kapitan, R., Carlson, M. D., Kuerbis, P. J., Clark, R. C., Melle, G. M., Sachse, T. P., & Walton, E. (1990). *Elementary school science for the 90s*. Andover, MA: The Network.

Madaus, G. (1987). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 82-21) (Eighty-seventh Yearbook of the National Society for the Study of Education). Chicago: The University of Chicago Press.

Malcolm, R., Burry, J. A. (1990). *What teachers need to know about assessing student achievement: Applications of the new NCME, AFT, and NEA standards*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA.

Marshall, H. H. (1988). Work or learning: Implications of classroom metaphors. *Educational Researcher*, 179-16.

Marshall, H. H. (1990). Beyond the workplace metaphor: The classroom as a learning setting. *Theory into Practice*, *29*(2), 94-101.

McCarthey, S. J., & Peterson, P. L. (1989). *Teacher roles: Weaving new patterns in classroom practice and school organization*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

McNeil, L. M. (1990). Empowering students: Beyond defensive teaching in social studies. In C. Emihovich (Ed.), *Locating learning: Ethnographic perspectives on classroom research* (pp. 117-140). Norwood, NJ: Ablex Publishing Corp.

Morine-Dershimer, G. (1989). Preservice teachers' conceptions of content and pedagogy: Measuring growth in reflective, pedagogical decision-making. *Journal of Teacher Education, 30*(5), 47-52.

Murphy, J. (1990a). Preparing school administrators for the 21st century: The reform agenda. In B. Mitchell & L. L. Cunningham (Eds.), *Educational leadership and changing contexts of families, communities, and schools*. Chicago: University of Chicago Press.

Murphy, J. (1990b). The educational reform movement of the 1980s: A comprehensive analysis. In J. Murphy (Ed.), *The reform of American public education in the 1980s: Perspectives and cases*. Berkeley: McCutchan.

Murphy, J. (1991). *Restructuring schools: Capturing and assessing the phenomena*. New York: Teachers College Press.

National Board for Professional Teaching Standards. (1991). *Toward high and vigorous standards for the teaching profession: Initial policies and perspectives of the National Board for Professional Teaching Standards* (3rd ed.). Detroit: The author.

National Committee on Science Education Standards and Assessment. (1993). *National Science education standards: An Enhanced Sampler*. Washington, D.C.: National Research Council.

Novak, J.D. (1990). Concept maps and vee diagrams, two metacognitive tools to facilitate meaningful learning. *Instructional Science 19, 29-52.*

Novak, J. D. & Gowin, D.B. (1984). *Learning how to learn.* New York: Cambridge University Press

Petrie, H. G. (1990). Reflections on the second wave of reform: Restructuring the teaching profession. In S. L. Jacobson & J. A. Conway (eds.), *Educational leadership in an age of reform.* New York: Longman.

Pittman, K. M. (1992). *Differences in the characteristics of extreme groups of expert science teachers: Implications for instructional practice.* (Unpublished Doctoral Dissertation, The University of Alabama, Tuscaloosa).

Resnick, L. (1987). *Education and learning to think.* Washington, DC: National Academy.

Salinger, G. L. (1991). The materials of physics instruction. *Physics Today, 44*(9), 39-45.

Schroder, H. M., Driver, M. S., & Streufert, S. (1967). *Human information processing: Individuals and groups functioning in complex social situations.* New York: Holt, Rinehart, & Winston.

Scriven, M. (1990). Can research based teacher evaluation be saved? *Journal of Personnel Evaluation, 4*(1), 19-32.

Scriven, M. (1991). Teacher evaluation models project memorandum. Kalamazoo: The Center for Research on Educational Accountability and Teacher Evaluation, Western Michigan University.

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction.* White Plains, NY: Addison-Wesley Longman, Inc.

Sparks-Langer, G. M., & Colton, A. B. (1991). Synthesis of research on teachers' reflective thinking. *Educational Leadership, 48*(6), 37-44.

Stufflebeam, D. L.,et al. (1990). *Center for Research on Educational Accountability and Teacher Evaluation (CREATE)* technical proposal funded by United States Office of Educational Research and Improvement (OERI) Awarded to Western Michigan University (Dan Stufflebeam, director) PR AWARD Number r117 Q00047, AFDA 84.117Q. Center for Evaluation: Kalamazoo,MI.

Turner, M. & Burry-Stock, J. (1993). *Expert science teaching evaluation model (ESTEEM); teaching practices inventory (TPI); A self-report analysis.* Paper presented at the Mid-South Educational Research Association, New Orleans, LA.

von Glasersfeld, E. (1988). *Cognition, construction of knowledg., and teaching.* Washington, DC: National Science Foundation.

Werner, H. (1957). The concept of development from a comparative and organismic point of view. In D. B. Harris (ed.), *The concept of development*. Minneapolis: University of Minnesota Press.

Wise, A. E. (1989). Professional teaching: A new paradigm for the management of education. In T. J. Sergiovanni, & J. H. Moore (Eds.), *Schooling for tomorrow: Directing reforms to issues that count*. Boston: Allyn & Bacon.

Yager, R.E. & Roy, R. (1993). STS: Most pervasive and most radical of reform approaches to "science" education. In R.E. Yager (Ed.) *What research says to the science teacher: vol.7.*

Yager, R. E. (1991). The constructivist learning model. *The Science Teacher, 51*, 52-57.